HOMEWORK #3
SOLUTIONS

Statistics 10

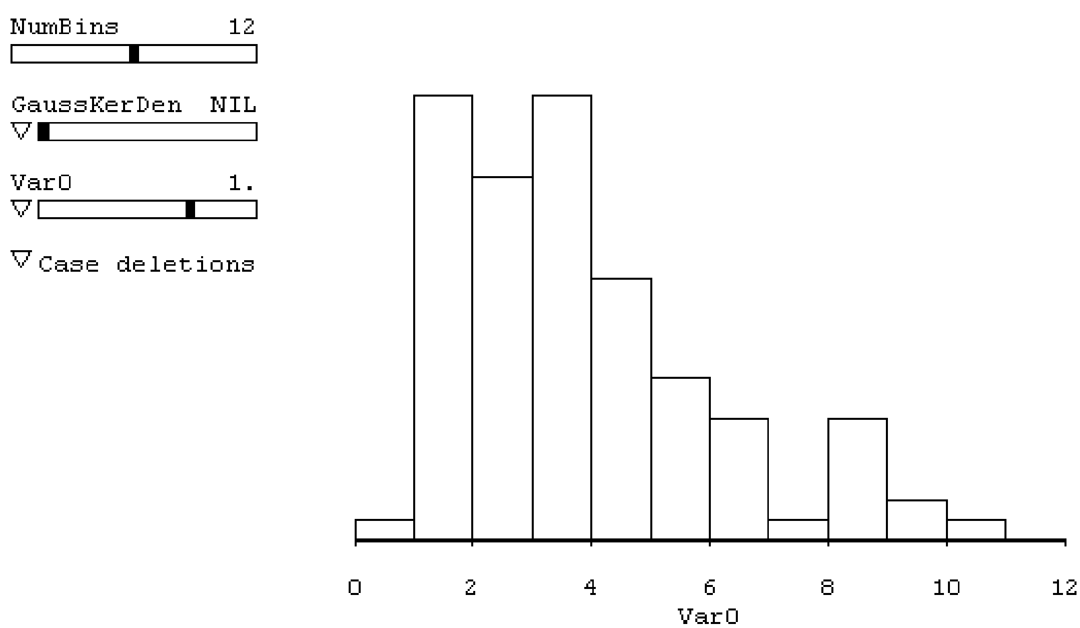
William Anderson,
Teaching Assistant

#3.1

The frequency table is:

Value	Observed Frequency
1	1
2	22
3	18
4	22
5	13
6	8
7	6
8	1
9	6
10	2
11	1

This can also be represented by a histogram:



The mean or average length of a word from the data is **4.4 characters**. The average length is found by summing up all of the observations, and then dividing by the number of observations. The mathematical notation is

$$Mean = \overline{X} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

where x_i denotes an individual observation and n is the number of observations. In this case, $n = 100$.

The mean or average length of a word from the frequency of occurrences is **4.4 characters**. This is called the expected value and is denoted $E(X)$. The expected value is calculated (for the discrete case) by summing up the values of the observations multiplied by the probability of the observation. The mathematical notation is

$$E(X) = \sum_{i=1}^{n} x_i p(x_i)$$

where x_i denotes an individual observation, $p(x_i)$ is the probability of the observation x_i, and n is the number of observations. In this case, the probability of the observation is just the observed frequency divided by the number of observations. For example, the probability of observing a word length of two is 22/100 or 0.220. Hence, we have:

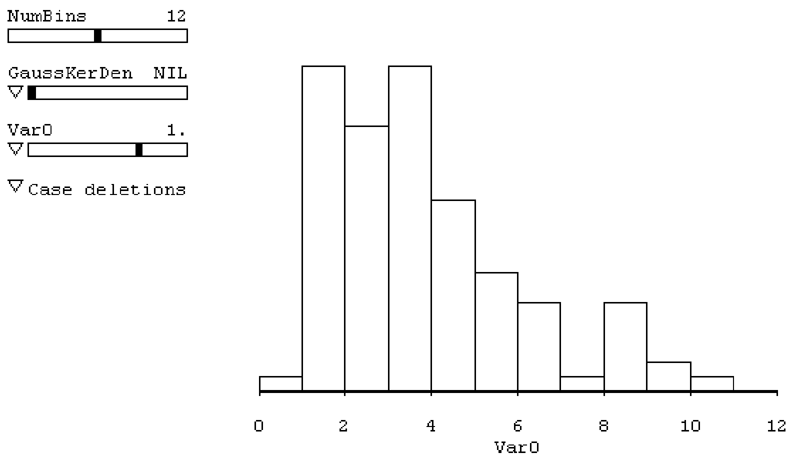
$$E(X) = 1\left(\frac{1}{100}\right) + 2\left(\frac{22}{100}\right) + 3\left(\frac{18}{100}\right) + 4\left(\frac{22}{100}\right) + 5\left(\frac{13}{100}\right) + 6\left(\frac{8}{100}\right) +$$
$$+ 7\left(\frac{6}{100}\right) + 8\left(\frac{1}{100}\right) + 9\left(\frac{6}{100}\right) + 10\left(\frac{2}{100}\right) + 11\left(\frac{1}{100}\right) = 4.4$$

The standard deviation of the length of a word is **2.2 characters**. This is found by summing up the squared differences of each observation and the mean, dividing by the number of observations (less one), and then taking the square root. The mathematical notation is

$$s = \sqrt{\frac{\sum_{i=1}^{n}\left(x_i - \overline{X}\right)^2}{n-1}}$$

where $\overline{X}$ is the mean of the observations.

Using the estimates derived above, we would expect the distribution of English words to follow the normal distribution (bell shaped curve) with a mean word length of 4.4 and a standard deviation of 2.2. We can sort of see this in the histogram:



The chance of randomly selecting a word of length 2 is

$$\Pr(X = 2) = \frac{22}{100} = 0.22$$

or **22%**.

#3.2

$X \sim N(3,4)$

The standard normal distribution is $Z \sim N(0,1)$, so the Z values are the number of standard deviations away from the mean. The normal distribution is symmetric about the mean.

For x=-5, we have

$$Z = \frac{X-\mu}{\sigma} = \frac{X-3}{4} = \frac{-5-3}{4} = -2,$$

or **–5 is 2 standard deviations away from the mean**.

For x=11,

$$Z = \frac{X-3}{4} = \frac{11-3}{4} = 2$$

or **11 is 2 standard deviations away from the mean**.

For x=5,

$$Z = \frac{X-3}{4} = \frac{5-3}{4} = 0.5$$

or **5 is 0.5 standard deviations away from the mean**.

For x=1.4,

$$Z = \frac{X-3}{4} = \frac{1.4-3}{4} = -0.4$$

or **1.4 is 0.4 standard deviations away from the mean**.

Finally, since the normal distribution is continuous,

$$\Pr(-3 < X < -1) = \Pr(-3 \le X \le -1)$$

because the probability of observing a single observation is 0. Note that this is not true for a discrete distribution. Hence we have

$$\Pr(-3 < X < -1) = \Pr\left(\frac{-3-3}{4} \le Z \le \frac{-1-3}{4}\right) = \Pr(-1.5 \le Z \le -1) =$$

$$= 0.1587 - 0.0668 = 0.0919$$

An alternative way to solve the problem is:

$$\Pr(-3 < X < -1) = 1 - \Pr\left((-3 < X < -1)^C\right) = 1 - \Pr(X \le -3) \bigcup \Pr(X \ge -1) =$$

$$= 1 - \Pr\left(Z \le \frac{-3-3}{4}\right) \bigcup \Pr\left(Z \ge \frac{-1-3}{4}\right) = 1 - \Pr(Z \le -1.5) \bigcup \Pr(Z \ge -1) =$$

$$= 1 - (0.5 - 0.4331) + (0.3413 + 0.5) = 1 - 0.9082 = 0.0919$$